Titanic Data Visualization

Using Spark with JAVA

Student:

PATIL Kunal

Teacher:

BROUSSARD Thomas

Course:

JAVA and UML Programming

Overview

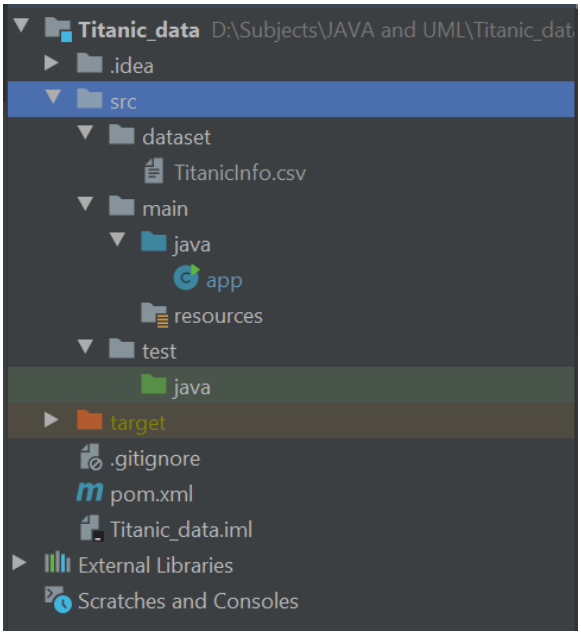
The project aims to create simplified and better view to understand of Titanic Dataset for the user. Removing unnecessary/Null data and processing correct data give data visualization is simple format.

Project: Titanic Data

Resources Used

Programming Language and Version	JAVA 11
IDE	IntelliJ IDEA 2020.1.2
Framework	Apache Spark 3.0.0
Plugin	Scala 2.12
Editor	Notepad++

Arborescence



Exploring dataset

Column Name	DataType	Value Range	Useful	Altered/Dropped
Survived	Integer	0-1	Yes	No
PClass	Integer	1-3	Yes	No
Name	String	A-Z	Partial	Altered (Title)
Sex	String	Female-Male	Yes	Altered (SexIndexer)
Age	Integer	0-100	Yes	Altered (AgeGroup)
Siblings/Spouses Aboard	Integer	0-10	No	Dropped
Parents/Children Aboard	Integer	0-10	No	Dropped
Fare	Float	0.0-500.00	Yes	No

Data Processing and Visualization

Total Survival Rate and Survival Percentage

Survived: 1

Not Survived: 0

```
+--------+-----+
|Survived|Count|
+--------+-----+
|       1|  342|
|       0|  545|
+--------+-----+
```

```
+--------+----------+
|Survived|Percentage|
+--------+----------+
|       1|     38.56|
|       0|     61.44|
+--------+----------+
```

Observation: Number of people survived are less than number of people died.

Number of people on ship based on Gender

```
+------+-----+
|   Sex|Count|
+------+-----+
|female|  314|
|  male|  573|
+------+-----+
```

```
+------+----------+
|   Sex|Percentage|
+------+----------+
|female|      35.4|
|  male|      64.6|
+------+----------+
```

Observation: More males were present on the ship than females.

Survival Rate and Percentage based on Gender

Survived	Sex	Count
1	female	233
0	female	81
1	male	109
0	male	464

Survived	Sex	Percentage
1	female	26.27
0	female	9.13
1	male	12.29
0	male	52.31

Observation: More number of males died on the ship compared to females.

Number of people on ship based on Passenger Class

Pclass	count(1)
1	216
3	487
2	184

Pclass	Percentage
1	24.35
3	54.9
2	20.74

Observation: Almost more than half of the population was boarded with 3rd Passenger class.

Survival Rate and Percentage based on Passenger Class

Survived	Pclass	Count
0	1	80
1	1	136
0	2	97
1	2	87
1	3	119
0	3	368

Survived	Pclass	Percentage
0	1	9.02
1	1	15.33
0	2	10.94
1	2	9.81
1	3	13.42
0	3	41.49

Observation: Almost 41% population died was from passenger class 3 whereas survival rate is more for population in class 1.

Survival Rate and Percentage for Each Class based on Gender

Survived	Pclass	Sex	Count
1	1	female	91
0	1	male	77
0	1	female	3
1	1	male	45
0	2	female	6
1	2	male	17
0	2	male	91
1	2	female	70
1	3	female	72
1	3	male	47
0	3	female	72
0	3	male	296

Survived	Pclass	Sex	Percentage
1	1	female	10.26
0	1	male	8.68
0	1	female	0.34
1	1	male	5.07
0	2	female	0.68
1	2	male	1.92
0	2	male	10.26
1	2	female	7.89
1	3	female	8.12
1	3	male	5.3
0	3	female	8.12
0	3	male	33.37

Observation: Almost 1/3rd population boarded with 3rd Passenger class who were males died. Whereas more Passenger class 1 Females managed to survive and again males from 2nd class died than females.

Number of people on ship based on Age Group

Child: 0 to 18

Adult: >18

AgeGroup	Count
Adult	721
Child	166

AgeGroup	Percentage
Adult	81.29
Child	18.71

Observation: 81% Adults were on the ship.

Survival Rate and Percentage based on Age Group

Survived	AgeGroup	Count
0	Adult	457
1	Child	78
1	Adult	264
0	Child	88

Survived	AgeGroup	Percentage
0	Adult	51.52
1	Child	8.79
1	Adult	29.76
0	Child	9.92

Observation: almost half of the population died who was Adult. Whereas almost half of the children from overall population of children managed to survive.

Survival Rate and Percentage For each gender based on Age group

Survived	Sex	AgeGroup	Count
0	female	Child	29
1	male	Adult	82
0	male	Adult	405
0	female	Adult	52
1	female	Adult	182
1	female	Child	51
1	male	Child	27
0	male	Child	59

Survived	Sex	AgeGroup	Percentage
0	female	Child	3.27
1	male	Adult	9.24
0	male	Adult	45.66
0	female	Adult	5.86
1	female	Adult	20.52
1	female	Child	5.75
1	male	Child	3.04
0	male	Child	6.65

Observation: 46% of overall population who died was male adults.

Survival Rate and Percentage by Passenger Class and Age Group

Survived	PClass	AgeGroup	Count
0	3	Adult	288
1	1	Child	14
0	1	Adult	78
1	3	Child	41
1	3	Adult	78
0	2	Child	6
0	1	Child	2
1	1	Adult	122
1	2	Adult	64
0	2	Adult	91
1	2	Child	23
0	3	Child	80

Survived	PClass	AgeGroup	Percentage
0	3	Adult	32.47
1	1	Child	1.58
0	1	Adult	8.79
1	3	Child	4.62
1	3	Adult	8.79
0	2	Child	0.68
0	1	Child	0.23
1	1	Adult	13.75
1	2	Adult	7.22
0	2	Adult	10.26
1	2	Child	2.59
0	3	Child	9.02

Observation: From overall population, more number of adult died who were from class 3.

Number of people on ship of different Titles

Title	Count
Don	1
Miss	182
Col	2
Rev	6
Lady	1
Master	40
Mme	1
Capt	1
Mr	513
Dr	7
th	1
Mrs	125
Sir	1
Jonkheer	1
Mlle	2
Major	2
Ms	1

Title	Percentage
Don	0.11
Miss	20.52
Col	0.23
Rev	0.68
Lady	0.11
Master	4.51
Mme	0.11
Capt	0.11
Mr	57.84
Dr	0.79
th	0.11
Mrs	14.09
Sir	0.11
Jonkheer	0.11
Mlle	0.23
Major	0.23
Ms	0.11

Observation: Major categories among the population was Married Males (more than half), Unmarried Females (1/5th of population) and 14% of married female population.

Amount of Fare based on Survival Rate

Survived	Amount
0	12103.678
1	16551.229

Amount of Fare based on Passenger Class

Sex	Amount
female	13966.663
male	14688.245

Amount of Fare based on Passenger Class

PClass	Amount
1	18177.412
2	3801.842
3	6675.654

Observation: Amount of fare collected from first class is more than twice of sum of amount collected for class 2 and class 3.

Prediction for survival rate using ML models

Accuracy of Decision Tree model : 0.8166666666666667

Confusion Matrix:

		Predicted class	
		P	N
Actual class	P	96.0	15.0
	N	18.0	51.0

Feature calculation for each columns:

Passenger Class: 0.1935417036561642

Age: 0.09942478970050685

FamilySize: 0.09733191173998515

Fare: 0.009857557150909526

indexSex: 0.5998440377524343

Conclusion:

Predicting survival rate from various parameters provided within the dataset. Visualize possible fields to understand given dataset more clearly.